HIV vs Supercomputing

Tanmoy Bhattacharya, T-2; Marcus Daniels, S. Gnanakaran, Bette Korber, T-6

Rapidly evolving viruses pose one of the major public health threats today. Among these, the Human Immunodeficiency Virus (HIV) that causes the Acquired Immune Deficiency Syndrome (AIDS) is particularly devastating, infecting 33 million people, with millions of AIDS-related deaths and new infections each year. Vaccines against such highly variable viruses have been unable to cope with the diversity of circulating strains: When a vaccine immunogen is presented to the human body, the elicited immune memory fails to recognize most other strains of the virus. This situation calls for both a thorough understanding of the plasticity of the virus in its war against the human host, and an intelligent design of vaccines that would provide lasting immunity against the virus. Starting with the establishment of a central sequence repository for the virus, to establishing that the virus has been circulating in humans since the early part of the 20th century [1], LANL has been at the forefront of such theoretical biology research and has contributed substantially to the field. We also developed ideas of artificial immunogens that better capture the observed diversity of HIV strains than any natural strain can do [2,3]. While preliminary results on these artificial immunogens are sufficiently promising [4] to move them to human trials, it is desirable to advance the field of vaccine design from such data-mining techniques to biological knowledge-based approaches.

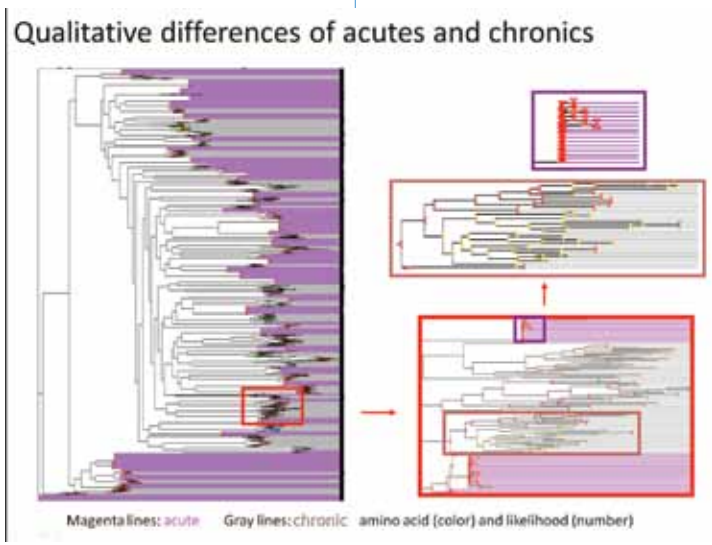

Fig. 1. Founder effects lead to entire clades sharing characteristics. If the sampling of the various clades is nonrandom, this leads to apparent correlations with traits of interest. True causal correlations, however, show up as correlations with changes.

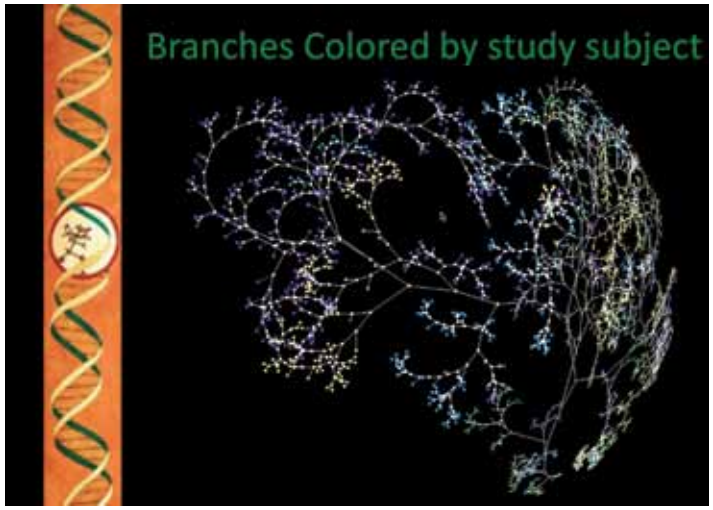

Fig. 2. A phylogeny of about 10,000 HIV sequences colored by the study subject that was used to implement phylogenetic correction on the observed correlation between genotype and phenotype.

The adaptive arm of the human immune system consists of three basic branches: 1) the Cytotoxic T Lymphocyte arm, which recognizes distinctive fragments of foreign proteins being manufactured in the body (i.e., viral proteins in infected cells) with very high specificity, 2) the Helper T Lymphocyte arm, which produces cytokines that orchestrate the immune response and have anti-viral activity, and 3) the B-cell or antibody arm, which recognizes distinctive shapes on the surface of fully folded proteins. Vast amounts of data on the interaction between HIV and all three arms of the human immune system are available, but the patterns of correlations are cryptic. Evolutionary systems are marked by long time scales, so that observed patterns in data can be due to correlations imposed by the initial historical emergence of a lineage of viruses, or *founder effects,* as well as due to biological interactions. In fact, not accounting for these effects leads to vastly erroneous statistical conclusions about the effect of the T-cell induced immunity on the evolution of the virus in populations [5]. But, whereas the sequence, i.e., the state of the virus, indeed depends on its evolutionary history, the changes that it undergoes are almost independent of changes in other lineages. Thus, true causal

correlations are also manifest in correlations with these changes (see Fig. 1), and in our work it was shown to detect effects that were validated experimentally. The separation of the two effects, i.e., a *phylogenetic correction*, thus needs access to these changes, and requires us to be able to statistically assess the genealogical relationships between the viruses and reconstruct the ancestral forms of the viruses.

Fortunately, evolution happens by the accumulation of random mutations, most of which are effectively neutral in that they do not affect the fitness of the virus to live and infect its hosts. The covariation of these mutations, then, carries a signal of shared history. This covariation can be used to construct a phylogenetic tree and an evolutionary model that leads to random changes, and, simultaneously, the ancestral forms of the virus are also reconstructed statistically. This reconstruction is technically challenging because the number of possible relationships grows factorially with the number of sequences sampled, and even heuristic searches fail to find reasonable models without extensive computations.

For example, a vaccine needs to prepare the body for fighting an incoming virus that can establish an infection in the healthy body. The virus that exists in a chronic patient, however, results from a long process of virus-host interaction and may be qualitatively different. The characterization of these differences is, however, a daunting task: the viral diversity in a chronic patient needs to be represented by at least three- to four-dozen sequences each, but to control for the phylogenetic effects, we need at least 200 to 300 patients who are infected with various subtypes of the virus. This means that one needs to fit together some 10,000 HIV sequences into a giant family tree of HIV viruses, and find those patterns that distinguish acute and chronic viruses. We therefore used Roadrunner to construct a phylogenetic tree of about 10,000 sequences from over 400 people (see Fig. 2) and are using it as a foundation to study the differences between acute and chronic sequences. This tree is currently being analyzed for pan-subtype signatures, as well as those that may be specific to individual subtypes.

We have since started analyzing data that directly measures the immunogenicity of viruses by collecting antibody-containing sera and viruses from the same patients. These sera are seen to cluster into groups with markedly different neutralization potencies [6]. The viral envelope (Env) glycoproteins, gp120 and gp41, are the main targets of antibody (Ab) neutralization. We therefore then looked at the Env sequences of viruses from the subset of patients who make potent sera that neutralize the activity of a vast panel of viruses, and compared them against those that make average or poor responses. Even though we did not know a priori whether the difference that we observed was due to host genetics, viral factors, or stochastic events, a preliminary analysis uncovered sites in the viral sequence where changes correlated with the induction of a good immune response. Interestingly, in a 3D X-ray structure of gp120 (see Fig. 3), these sites clustered around the part of the virus that binds to the human CCR5-coreceptor, an interaction that mediates viral entry into cells, and a part of the viral envelope that had long been suspected to be involved in the induction of beneficial antibodies. In the current analysis, however, we used a small number of sera. This analysis is being extended to a much larger panel.

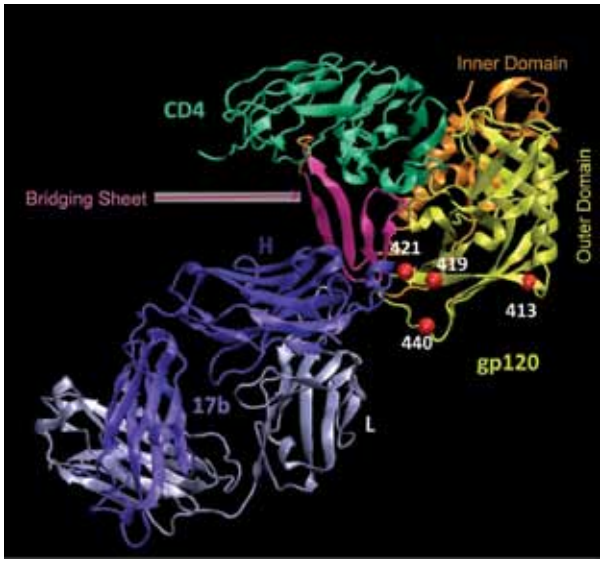

Fig. 3. Sites where mutations correlate with the induction of broadly neutralizing activity cluster in the CD4-inducible region of the HIV envelope glycoprotein gp120. When this protein (yellow, red, and orange) interacts with the cell surface molecule CD4 (green), a conformational change opens up the CD4-inducible region where our signature sites (red balls) are found. The X-ray structure (PDB code: 1RZK) also shows an antibody (blue) that binds to this region.

The advent of petaflop-scale computing, exemplified by Roadrunner, is coming to the rescue, and in the near future we expect to see a fully detailed phylogenetic analysis of such problems. Such computational techniques, complemented with our advance in experimental methods and theoretical understanding, we hope will usher in a new era that finally stops this deadly epidemic.

For more information contact Tanmoy Bhattacharya at tanmoy@lanl.gov.

[1] B. Korber et al., *Science* **288,** 1789 (2000).
[2] B. Gaschen et al., *Science* **296,** 2354 (2002).
[3] W. Fisher et al., *Nature Medicine* **13,** 100 (2007).
[4] S. Santra et al., *Proc. Natl. Acad. Sci. Unit. States Am.* **105,** 10489 (2008).
[5] T. Bhattacharya et al., *Science* **315,** 1583 (2007).
[6] N.A. Doria-Rose et al., *J. Virol.,* in press (2009).

Funding Acknowledgments
LANL Directed Research and Development Program